A new nearest neighbor approach for imbalanced learning

Presenter: Yuxuan Li
Supervisor: Dr. Jenny Zhang
Outline

- Imbalance problem in classification
- Approaches
  - SMOTE
  - $k$ Exemplar-based Nearest Neighbor
- Experimental results
Imbalanced Problem

- The number of majority instances heavily exceed the other class.

- Example:
  - The identification of oil spills in satellite radar images.
  - The detection of software defeats in large system.
  - The diagnoses of rare medical conditions such as thyroid diseases.

- The minority instances (positive) are critical in these domains.
Imbalanced Problem

- Traditional classifiers can suffer from imbalanced class distribution.
- Tend to predict the most probable majority class.
- Positive instances are submerged by negatives.

Fig. 1. An artificial imbalance classification problem
Current approaches

- SMOTE is a popular approach to address the imbalanced problem.
- SMOTE: Over-sampling the minority instances, to make the dataset more balanced.
- Other approaches: SHRINK, MetaCost
Our approach

- $k$ Exemplar-based Nearest Neighbor ($k$ENN)
- Introduce a training process for nearest neighbor.
- Identify the Pivot Positive Instances (PPI) in the training stage, by using pessimistic estimate.

**Fig. 2.** The Voronoi diagram for the subspace of subconcept P3 of Fig. 1.
Algorithm 1 Compute the set of positive pivot points

Input:  
a) Training set $T$ ($|T|$ is number of instances in $T$); b) confidence level $c$.

Output: The set of pivot positive instances $P$ (with radius $r$ for each Gaussian ball)

1: $\delta \leftarrow$ FP rate threshold by Equation (1) from $c$, $|T|$, and prior negative frequency 
2: $P \leftarrow \phi$
3: for each positive instance $x \in T$ do
4: $G \leftarrow$ neighbors of $x$ in increasing order of distance to $x$
5: for $k = 1$ to $|G|$ do
6: if $G[k]$ is a positive instance then
7: break \{;; $G[k]$ is the nearest positive neighbor of $x$\}
8: end if
9: end for
10: $r \leftarrow$ distance($x$, $G[k]$)
11: $f \leftarrow \frac{k-1}{k+1}$ \{;; Gaussian ball $B(x, r)$ has $k+1$ instances and $(k+1-2)$ FPs\}
12: $p \leftarrow$ the FP rate by Equation (1) from $c$, $k$ and $f$
13: if $p \leq \delta$ then
14: $P \leftarrow P \cup \{x\}$ \{;; $x$ is a pivot positive instance, and $P$ is the output\}
15: end if
16: end for

\[
f + z^2/2N + z\sqrt{f(1-f)/N} + z^2/4N^2 \over 1 + z^2/N \]
Datasets:
We test on 12 imbalanced dataset from the UCI database and other disciplines.

Metric:
Area Under ROC Curve (AUC)

Overall accuracy is not a suitable measurement for imbalanced dataset.

Example:
A software defect problem. Total 10 instances, only one has defect.
A trivial classifier predict every instance as negative.
Accuracy = Correct predict / Total = 9 / 10 = 90%
ROC curve

- ROC curve is a class distribution independent metric.
- ROC graphs have been increasingly used in the data mining, especially for imbalanced domains.

The area under the curve (AUC) is used to compare classifier performance.

A perfect classifier has AUC of 1.
## Comparisons

<table>
<thead>
<tr>
<th>Dataset</th>
<th>3ENN</th>
<th>Naive</th>
<th>3NN</th>
<th>3NNSmt+</th>
<th>3NNMeta</th>
<th>C4.5</th>
<th>C4.5Smt+</th>
<th>C4.5Meta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oil</td>
<td>0.811</td>
<td>0.788</td>
<td>0.796</td>
<td>0.797</td>
<td>0.772</td>
<td>0.685</td>
<td>0.771</td>
<td>0.764</td>
</tr>
<tr>
<td>Hypo-thyroid</td>
<td>0.846</td>
<td>0.831</td>
<td>0.849</td>
<td>0.901</td>
<td>0.846</td>
<td>0.924</td>
<td>0.948</td>
<td>0.937</td>
</tr>
<tr>
<td>PC1</td>
<td>0.806</td>
<td>0.786</td>
<td>0.756</td>
<td>0.755</td>
<td>0.796</td>
<td>0.789</td>
<td>0.728</td>
<td>0.76</td>
</tr>
<tr>
<td>Glass</td>
<td>0.749</td>
<td>0.623</td>
<td>0.645</td>
<td>0.707</td>
<td>0.659</td>
<td>0.696</td>
<td>0.69</td>
<td>0.754</td>
</tr>
<tr>
<td>Satimage</td>
<td>0.925</td>
<td>0.839</td>
<td>0.918</td>
<td>0.902</td>
<td>0.928</td>
<td>0.767</td>
<td>0.796</td>
<td>0.765</td>
</tr>
<tr>
<td>CM1</td>
<td>0.681</td>
<td>0.606</td>
<td>0.637</td>
<td>0.666</td>
<td>0.625</td>
<td>0.607</td>
<td>0.666</td>
<td>0.668</td>
</tr>
<tr>
<td>New-thyroid</td>
<td>0.99</td>
<td>0.945</td>
<td>0.939</td>
<td>0.972</td>
<td>0.962</td>
<td>0.927</td>
<td>0.935</td>
<td>0.931</td>
</tr>
<tr>
<td>KC1</td>
<td>0.794</td>
<td>0.732</td>
<td>0.759</td>
<td>0.756</td>
<td>0.779</td>
<td>0.64</td>
<td>0.709</td>
<td>0.695</td>
</tr>
<tr>
<td>SPECT_F</td>
<td>0.767</td>
<td>0.728</td>
<td>0.72</td>
<td>0.725</td>
<td>0.735</td>
<td>0.626</td>
<td>0.724</td>
<td>0.643</td>
</tr>
<tr>
<td>Hepatitis</td>
<td>0.783</td>
<td>0.71</td>
<td>0.758</td>
<td>0.772</td>
<td>0.744</td>
<td>0.753</td>
<td>0.713</td>
<td>0.745</td>
</tr>
<tr>
<td>Vehicle</td>
<td>0.952</td>
<td>0.945</td>
<td>0.969</td>
<td>0.942</td>
<td>0.956</td>
<td>0.921</td>
<td>0.926</td>
<td>0.929</td>
</tr>
<tr>
<td>German</td>
<td>0.714</td>
<td>0.677</td>
<td>0.69</td>
<td>0.686</td>
<td>0.705</td>
<td>0.608</td>
<td>0.649</td>
<td>0.606</td>
</tr>
<tr>
<td>Average</td>
<td>0.818</td>
<td>0.768</td>
<td>0.786</td>
<td>0.798</td>
<td>0.792</td>
<td>0.745</td>
<td>0.771</td>
<td>0.766</td>
</tr>
</tbody>
</table>
Fig. 3. ROC curves with convex hull on two datasets. The x-axis is the FP rate and the y-axis is the TP rate. Points on the convex hull are highlighted with a large circle.
Summer Project Summary

- A camera-ready version paper for the *15th Pacific-Asia Conference on Knowledge Discovery and Data Mining*

- Extend to a Journal paper.
Thank you!